EXHIBIT G

Page 204

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

_____

RICHARD KADREY, ET AL.,   )

    INDIVIDUAL AND REPRESENTATIVE   )

                         PLAINTIFFS,) LEAD CASE NO.

            v.   ) 3:23-cv-03417-VC

META PLATFORMS, INC.,   )

                    DEFENDANT.)

_____)

  * * * H I G H L Y   C O N F I D E N T I A L   * * *

* * * A T T O R N E Y S '   E Y E S   O N L Y   * * *

VIDEO-RECORDED 30(B)(6) DEPOSITION OF

MICHAEL CLARK

VOLUME III

WEDNESDAY, NOVEMBER 20, 2024

SAN FRANCISCO, CALIFORNIA

9:38 A.M. PST

REPORTED BY AUDRA E. CRAMER, CSR NO. 9901

_____

DIGITAL EVIDENCE GROUP

1730 M Street, NW, Suite 812

Washington, D.C. 20036

(202) 232-0646

Page 312

```
 1   Hugo Touvron; Soumya, S-o-u-m-y-a, Batra; Todor,

 2   T-o-d-o-r, M-i-h-a-y-l-o-v; Sharan Narang,

 3   S-h-a-r-a-n, N-a-r-a-n-g; and Thomas Scialom,

 4   T-h-o-m-a-s, S-c-i-a-l-o-m.

 5        Q.    Okay.  We're going to get into the

 6   contents of this document.

 7             Did you review this document in

 8   preparation?  Are you familiar with it?

 9             MS. HARTNETT:  Object to the form.

10             THE WITNESS:  Let me read it real quick

11   just to see if I did.

12             Okay.

13   BY MS. POUEYMIROU:

14        Q.    Okay.

15        A.    I don't remember -- to answer your

16   question, I don't remember seeing this in my

17   preparation.

18        Q.    Okay.  And before this you were talking

19   about a script that Bashlykov made to remove

20   copyright information from LibGen; correct?

21             MS. HARTNETT:  Object to the form.

22             THE WITNESS:  To be specific, the data
```

Page 313

```
 1   that Bashlykov was applying the script was to

 2   actual data and not to metadata, which was

 3   separate.

 4   BY MS. POUEYMIROU:

 5       Q.   Okay.

 6       A.   It was inside of the training data or

 7   proposed training data itself.

 8       Q.   And was it inside the data because it

 9   was -- because copyright information is inside a

10   book?

11           MS. HARTNETT:  Objection to the form of

12   the question.

13           THE WITNESS:  The data that was removed

14   included the word "copyright," the word

15   "acknowledgments," and matches upon phrases and

16   lines in certain parts of the book that did

17   that.  It was removed because of the technical

18   risk that was introduced to the model by keeping

19   that data in.

20   BY MS. POUEYMIROU:

21       Q.   So the removal of that copyright

22   information facilitated training of the model?
```

Page 314

 1   MS. HARTNETT: Objection to the form.

 2   THE WITNESS: The removal of that data

 3   kept there from being additional noise

 4   introduced from repetitive phrases and from

 5   things that would introduce other noise or

 6   problems into the model itself.

 7   BY MS. POUEYMIROU:

 8   Q.   And would you say that removal of that

 9   facilitates training?

10   A.   Removal of that data is one step in the

11   data cleanup phase, which is one of many pieces

12   to facilitate training.

13   Q.   Okay. So in the opening of this

14   document, you have Sergey Edunov and Melanie

15   Kambadur noted.

16   And what I was hoping you could explain

17   is what is meant by this specific language: "We

18   need to make sure source metadata does not cause

19   problems in training, stability, performance or

20   leakage of sensitive info. We have experimental

21   data now showing basically no negative effects

22   for the v1 unhashed source metadata, but since

Page 315

```
 1    this is small-scale, noisy evals, it's

 2    unfortunately not 100 percent conclusive.

 3    Intuitively, adding 5 to 10 percent to token

 4    size could be a risk."

 5            And this is the sentence in particular

 6    I'm interested in:  "Also, unhashed source

 7    metadata is not acceptable from an IP-brand

 8    safety leakage perspective IMO.  We have to be

 9    basically 100 percent confident no one can

10    extract this data, and based on previous

11    memorization work, I do not think we can be."

12            So can you walk me through what they're

13    talking about.

14            MS. HARTNETT:  Object to the form.

15            THE WITNESS:  I was trying to find the

16    exact line.

17            Was that the part from Melanie, or was

18    that the --

19    BY MS. POUEYMIROU:

20       Q.   It looks like -- I mean, it's

21    "Melanie:" and then @sergeyedunov?

22            MS. HARTNETT:  I think she's reading on
```

Page 344

```
 1                    (At 1:39 P.M. PST the

 2               deposition of MICHAEL CLARK was

 3               reconvened.)

 4          THE VIDEOGRAPHER:  We are now on the

 5    record at 1:39.

 6

 7               EXAMINATION (CONTINUED)

 8    BY MS. POUEYMIROU:

 9       Q.   Good afternoon, Mr. Clark.

10       A.   Good afternoon.

11       Q.   So I wanted to talk about a different

12    dataset now.

13       A.   Okay.

14       Q.   I mean, we have been talking about it,

15    but LibGen.

16               And so last week we looked at a

17    document that showed it being downloaded October

18    of 2022.

19               And your testimony today is you don't

20    know who did that download -- the initial

21    download of LibGen; is that correct?

22       A.   Just to --
```

Page 349

 1   Q.   This says, "Decided to go with direct

 2   file upload without using torrents for the

 3   following reasons: Using torrents would entail

 4   seeding the files, sharing the content outside.

 5   This could be legally not okay."

 6   So how did Meta obtain and copy LibGen?

 7   MS. HARTNETT: Object to the form.

 8   THE WITNESS: So I spoke with

 9   Bashlykov. We used libtorrent and the torrent

10   protocol for downloading. We had tried to

11   download a few different ways. One of them

12   included doing it via direct downloads where

13   there were individually hosted files in

14   different places and saw many issues with that

15   where files wouldn't exist, it was too slow or

16   would error out.

17   Also tried -- well, actually, the other

18   only other option was downloading via torrent,

19   and --

20   BY MS. POUEYMIROU:

21   Q.   And that's what you did?

22   A.   That's what we did. And the library

Page 350

```
 1   that we used using, LibGen -- sorry --

 2   libtorrent for downloading LibGen, Bashlykov

 3   modified the config setting so that the smallest

 4   amount of seeding possible could occur.

 5       Q.    What does that mean?

 6       A.    When you use a torrent protocol, part

 7   of the configuration on how torrents work is you

 8   can only download as long as you offer to

 9   participate in the torrent network in some way,

10   and seeding means that you have to open up some

11   amount of sharing of the torrent file while

12   you're downloading.

13       Q.    So why did Meta decide to minimize

14   seeding the files?

15            Is it because it was concerned that

16   that was legally not okay, to quote this

17   document?

18            MS. HARTNETT:  Objection to the form.

19            And to the extent it calls for any

20   attorney-client privilege, please don't reveal

21   that.

22            THE WITNESS:  Yeah, I do not know what
```